融合跨模态语义检索与条件扩散生成的三维手势合成方法

王欣艺[1)] 刘世光[1)] 杨照萌[2)]

[1)](天津大学智能与计算学部 天津 300354)

[2)](北京师范大学艺术与传媒学院 北京 100875)

摘 要 高质量虚拟人物手势生成在影视动画制作、数字人合成、手势舞编排等领域具有重要的应用价值。然而，目前基于生成模型的方法普遍面临多模态语义关联性弱、可控性不足等技术难点。本研究提出一种融合跨模态语义检索与条件扩散生成的三维手势合成框架。本文首先构建包含文本描述、音频特征与三维动作参数的多模态手势数据库；其次，本文提出分层对比语义检索机制，通过低级特征检索模块与高级语义检索模块的双阶段架构，并结合动量蒸馏策略显著提升跨模态检索精度与噪声鲁棒性；接着构建了基于多级注意力引导的扩散生成模型，通过跨模态注意力层与命令自适应归一化(PAAN)层，实现文本提示驱动的风格控制与未匹配动作的手势生成；然后针对动作序列衔接问题，提出加权融合策略与扩散模型的渐进式去噪机制，有效解决动作间不自然过渡问题，最终实现高质量的三维手势生成。在 BEAT 数据集上的定量实验说明，本方法在弗雷歇手势距离等指标上优于对比的手势生成方法。弗雷歇手势距离（FGD）为 9.160，动作多样性（Div）达 12.77，语义匹配度（SRGR）为 0.205。定性分析与用户评估结果显示，本方法生成手势的人类相似度评分较 LivelySpeaker 提升 24.9%，语义适当性评分较 GestureDiffuCLIP 提升 32.7%，在语义表达准确性与动作自然度上展现出更强的动作表现力。

关键词 手势生成;分层检索;对比学习;扩散模型；动量蒸馏

中图法分类号 TP391

Gesture Synthesis Method Integrating Cross-Modal Semantic Retrieval and Diffusion Model

WANG Xin-Yi[1)] LIU Shi-Guang[1)] YANG Zhao-Meng[2)]

[1)](College of Intelligence and Computing, Tianjin University, Tianjin 300354)

[2)](School of Art & Communication, Beijing Normal University, Beijing 100875)

Abstract Gestures are a crucial form of expression in human communication, particularly in domains such as virtual human interaction. In this field, speech-driven virtual human gesture generation technology acts as a core enabler to enhance the interactivity and immersion of virtual characters. High-quality virtual character gesture generation holds significant application value in film and animation production. To address existing technical bottlenecks of weak semantic association and insufficient control-lability in current generation methods, we propose the gesture synthesis method integrating cross-modal semantic retrieval and diffusion model. First, we construct a multi-modal gesture database containing text descriptions, audio features, and 3D motion parameters. Second, we innovatively design the Hierarchical Contrastive Semantic Retrieval Mechanism that adopts a dual stage architecture combining primary feature retrieval and semantic retrieval, significantly enhancing cross-modal alignment accuracy. Meanwhile, to further enhance the noise robustness of the retrieval mechanism and avoid retrieval errors induced by audio noise and text ambiguity, we introduce the momentum distillation

本课题得到国家自然科学基金项目(No. 62072328)资助。王欣艺，硕士研究生，主要研究领域为虚拟人物手势生成。刘世光（**通信作者**），博士，教授，中国计算机学会杰出会员，主要研究领域为计算图形学、计算机动画、虚拟现实。杨照萌，本科生，主要研究领域为舞蹈学、舞蹈教育。

strategy. Through model distillation, we strengthen the feature extraction capability and generalization ability of the retrieval model, which substantially boosts the accuracy and stability of cross-modal alignment. Furthermore, we develop the diffusion generation model that achieves text-prompt-driven style control and unmatched gesture generation through cross-modal attentions and Prompt Adaptive Attention Normalization (PAAN) layers. To address motion sequence transition issues, we propose a weighted fusion strategy and progressive denoising mechanism for diffusion models, effectively resolving unnatural gesture transitions. To comprehensively verify the effectiveness and superiority of the proposed method, we conduct systematic quantitative evaluations on the publicly available BEAT dataset. As a widely adopted benchmark dataset in the field of virtual human gesture generation, it can objectively reflect the generation performance of models. Quantitative evaluations on the BEAT dataset demonstrate the effectiveness of our method in key metrics: our method achieves a Fréchet Gesture Distance (FGD) of 9.160, the Diversity (Div) of generated gestures reaches 12.77, and the Semantic-Relevance Gesture Recall (SRGR) is 0.205. In qualitative experiments and user evaluations, our method achieves significant improvements in semantic consistency and motion naturalness of generated gestures: the human similarity score of our generated gestures is 24.9% higher than that of LivelySpeaker, and the semantic appropriateness score is 32.7% higher than that of GestureDiffuCLIP, demonstrating stronger expressive power in terms of semantic expression accuracy and motion naturalness. This performance gain stems from the collaborative design of multiple modules in our study, which effectively breaks through the bottlenecks of existing generation models in cross-modal fusion and motion controllability. The Retrieval Mechanism provides precise support for cross-modal information alignment and addresses the disconnection between semantic and motion features; the diffusion generation model guided by multi-level attention balances the stylistic diversity and semantic accuracy of gesture generation, and the introduction of the PAAN layers enables the model to flexibly respond to the stylistic demands of different text prompts; meanwhile, the motion transition optimization strategy further enhances the coherence of gesture sequences and avoids unnatural issues. Our method can accurately capture the core features of gestures corresponding to long-tail semantic content, and exhibits excellent performance in semantic matching accuracy, motion naturalness, which significantly enhances the reliability of gesture generation in complex scenarios.

Key words gesture generation; hierarchical retrieval; contrastive learning; diffusion models; momentum distillation

1 引言

在日常生活中，无论在公开演讲或是对话场景中，手势都起着至关重要的作用。手势是人类在谈话间自发进行且富有风格的手臂动作，它使语言更加生动，使表达充满活力，还能够传达人类表达的隐含思想[1]。除此以外，手势还有助于强调对话的情感，并能将口头描述的内容以可视化的形式展现出来。因此，在大型游戏或动画制作中，创作者往往寻求更丰富、更具个性、更自然逼真的角色表演，就需要虚拟人物表现出真实的手部动作以提高娱乐性和用户沉浸性。

但是，现代游戏与动作制作中角色动画越发复杂，这种复杂性不仅体现在视觉呈现的难度上，更体现在角色与玩家的互动方式中。单靠传统的运动捕捉与动画合成技术，难以满足现代场景中角色手势生成的真实性需求。基于此，语音驱动的虚拟人物自动手势生成方法受到越来越多的关注。使用自动生成手势的方式不仅能高效赋予虚拟角色生命力与真实性，还极大地提升了在互动场景中的表现力和沉浸感，但生成丰富的虚拟人物动作是一个复杂的任务。尤其是在当前技术条件下，完全基于生成模型合成复杂的手势动作时，仍然存在可控性差、生成质量低等问题。这是由于在人类自然手势分布中，高频常见手势（如挥手、手臂上下挥舞）占据分布的主体，而大量具有特定语义的低频手势（如手势模拟特定的数字等动作）形成分布的长尾，这些动作的稀疏性使得生成模型捕捉具有强 6 义的手势非常困难。虽然目前最先进的基于扩散模型的方法表现出一定语义与手势的相关性，但是仍然集中在一些简单的手势动作上，难以完成特定语

义的动作(如手指模拟数字，手臂拒绝动作等)。

为了解决以上问题，本文基于人类手势表达中复用-生成认知[2-5]，本质是复用已有熟悉手势，然后再动态微调与衔接动作，例如说“第一点”时，会复用“伸手指”的熟悉手势，再自然过渡到下一个动作。本文设计了基于检索式与生成式的混合框架，并构建了包含文本描述、音频特征与三维动作参数的多模态手势数据库。接着，本文提出了分层对比语义检索机制，设计实现了帧级检索与片段级语义检索的双阶段架构，并结合动量蒸馏策略，显著提升跨模态检索精度与噪声鲁棒性。本文还构建了基于多级注意力引导的扩散生成模型，通过跨模态注意力层与命令自适应归一化层，实现文本提示驱动的风格控制与未匹配动作的手势生成。针对动作序列衔接问题，提出加权融合策略与扩散模型加噪去噪机制，保证动作间的自然衔接。因此，检索模块直接从高质量动作库中匹配长尾高语义手势，无需生成模型无样本学习，从源头保证语义准确性，生成模块则负责过渡动作、节奏同步动作，符合基本的复用-生成的人类认知理论，保证流畅性与多样性。最终，本文方法不仅可以高效地从语义动作库中检索出与音频语义匹配的动作，还能利用扩散模型生成手势库中未匹配的动作以及之间的过渡与节奏，从而实现高质量的手势动作生成。

2 相关工作

在早期阶段，通过运动捕捉技术为虚拟角色生成手部动作，被广泛应用于电影和游戏行业。然而，这种技术成本高昂且耗时巨大。因此，生成与语音同步的手势任务逐渐成为研究热点，本节主要对现有的手势生成相关工作进行讨论。

2.1 基于规则驱动的手势生成

基于规则的手势创建方法优势在于能够生成与语音同步且质量上乘的手势动作。这种技术基于预设的动作库、手部动画或者精心设计的手势生成规则，因此其动作效果上更自然流畅。Wagner 等人[6]率先提出一种动画对话框架，通过对话规划器生成对话文本和语调，再结合面部表情生成器、手势生成器和动画系统进行统筹协调，实现与虚拟角色的互动和交互。Kopp 等人[7]使用 XML 语言来描述话语的外在形式，并结合非均匀三次样条曲线来构建满足所有位置和速度约束的复合曲线，以生成平滑的轨迹并再现人类运动的对称钟形速度轮廓。之后通过信念-欲望-意图（Belief–Desire–Intention，BDI）架构，设计了包含情绪，动作和行为的格式标准，加强在特定情境下执行包括头部运动和人类手势的复杂任务的能力。

虽然基于人类专家精心构建数据库的方式能细致调整映射规则，增强其可解释性和与可控性。但是，这种基于规则的方法需要大量特定领域的动作标注，同时需要专业词汇与手势动作之间的精确对应，这使得完全依赖数据库匹配的方法，在鲁棒性和灵活性上存在不足。此外，这类方法的动作效果受限于数据集的质量，意味着规则型方法只能在一定程度上为特定的语音或场景生成有限的合适手势。

2.2 基于数据驱动的手势生成

早期的数据驱动方法通过统计模型学习语音与手势的映射。基于统计模型的方法[8-9]通过结合预定义的动画单元，再通过构建数据的隐含规则来生成手势。这类方法需要在手势数据上预先计算概率分布，并依据语音输入从分布中采样以生成手势。与基于规则的方法相比，统计模型展现出更高的适应性，能够捕捉手势在对话中的不确定性。然而，统计模型在生成多样性手势方面的能力受限。

还有研究者们引入了生成对抗网络（Generative Adversarial Networks，GAN）[10]、变分自编码器（Variational Autoencoders，VAE）[11]和流模型（Flow-based Model）[12]来生成手势。然而基于 GAN 的方法存在模型崩溃与训练困难问题，基于 VAE 和 Flow 则难以在生成明确语义手势与多样性手势间平衡[13]。为克服这些缺点，扩散模型[14]被引入手势生成领域，用于生成运动序列[15-19]。Rombach 等人[15]提出使用扩散模型生成任意手势长度序列的方法，从而提升了语义效果与计算效率。Zhu 等人[17]提出带有退火噪声采样策略的扩散模型生成手势。Yang 等人[20]在手势扩散管道中引入了注意力机制，以生成与语音相匹配的手势。研究者们还提出了用户通过自然语言描述的手势风格[21]以提高可控性的方法。Yang 等人[18]在手势生成时直接将 CLIP（Contrastive Language-lmage Pretraining）编码的风格提示作为条件，以实现生成具有风格的手势，但生成效果仍然不明显。Ao 等人[16]通过自适应实例归一化（Adaptive Instance Normalization，AdaIN）层注入风格特征以支持多模态的风格提示，虽然提示效果优于之前的方法，但在解决抖动问题上仍表现不足。尽管上述方法通过扩散建模增强了语义感知，但是其手势生成能力与可控性仍然有提升的空间。

3 本文方法

本文方法旨在解决现有方法在语义关联性、可控性方面的不足。本文模型根据用户输入的音频信息，自动生成语义匹配、风格可控且自然流畅的三维手势序列。模型的必需输入为音频 A，本文系统会自动将输入音频转录为对应的文本描述 T。此外，用户可以选择性地提供初始手势 S（默认为系统自动生成的初始姿态）用于引导生成起点，以及可选的输入文本命令进行样式提示 P，用于控制生成手势的风格特征（如情感或特定动作）。模型的最终输出为三维手势序列文件（.bvh 格式），可直接渲染或利用蒙皮技术进一步细化角色动画效果，其中三维手势序列定义为 $G=[g_0,...,g_t]$，其中 g_t 表示 t 时刻的手势状态。

$$G = Model(A,T,S,P) \tag{1}$$

整体方法的框架流程如图 1 所示，本研究提出了融合跨模态语义检索与条件扩散生成的三维手势合成方法，可分为三个阶段：

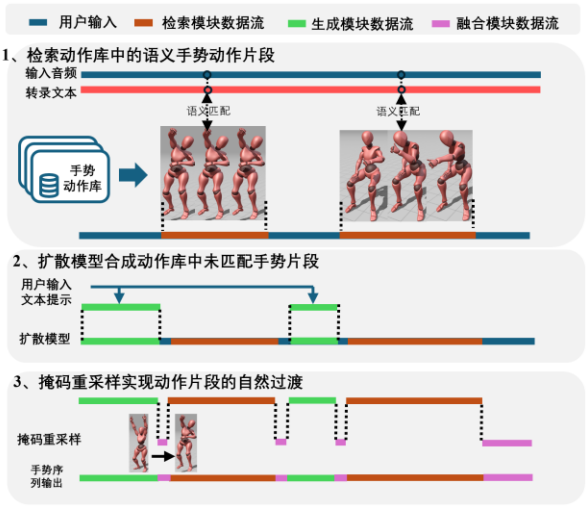

图 1 本文提出的手势生成框架

(1) 本文首先构建了包含文本描述、音频特征与三维动作参数的多模态数据库。用户只需输入音频，系统便可自动生成转录文本，并检索动作库中与音频内容相匹配的语义手势片段。需要注意的是，图 1 中手势动作库与多模态数据库并非一致，多模态数据库包含文本描述、音频特征、三维动作参数三类数据，手势动作库存储动作参数用于检索。图 1 中检索动作库中的语义手势动作流程包含图 2 （构建多模态数据库）与图 3 （分层对比检索语义手势）两部分的内容。

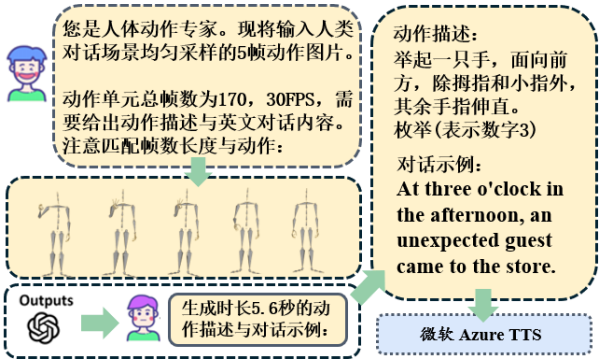

图 2 构建多模态数据库的方法流程示意图

(2) 通过多级注意力引导的扩散模型生成未匹配动作库中的手势，用户可通过输入文本提示来指定特定风格，其中图 1 中扩散模型合成动作库中未匹配手势流程对应图 4（扩散模型生成手势流程）相应的文本内容。

(3) 通过基于加权融合策略与扩散模型精细化动作进行手势序列的衔接，将检索到的语义手势与生成模型合成的手势集成在一起，最终输出自然连贯的三维手势序列。

3.1 动作库的构建

本文构建多模态数据库的方法流程如图 2 所示，动作库的构建使用了 MOCCA[22]语义手势数据集，每个动作元都包含一组手势动作与对应信息。为了建立跨模态语义关联空间，本研究将手势数据集中每个语义手势单元均匀采样 5 帧关键图像，构建具有时序代表性的视觉提示序列。采用 GPT-4（Generative Pre-trained Transformer 4）[23]作为多模态决策引擎，在视觉模态输入 5 帧均匀采样的 RGB 图像序列，覆盖手势起始、过渡与终止阶段。文本模态输入当前手势的数据描述（包含语义标签、运动总帧数等上下文信息）。多模态大模型最终输出动作库中每个手势单元的动作描述和样例文本，完整标注流程以人工判断为核心，GPT-4 仅负责生成初始候选描述，人工筛选后的文本描述通过由微软 Azure TTS（Text-to-Speech）工具[24] 合成音频信息。在 GPT-4 输入手势单元信息时，会同时提供该手势的总帧数与帧率，控制句子长度生成时长匹配的文本，另一方面将生成的文本输入 Azure TTS 时，会调整设置语速参数，并通过 TTS 的时长预览，若合成音频时长偏离，则微调语速直至时长对齐。若音频朗读后出现语义歧义或时长不一致，则返回人工环节修正文本表述，直到音频语义与动作语义完全匹配，才能确定为标注文本。最终构建包含文本描述、音频特征与三维动作参数的多模态手势数据库。

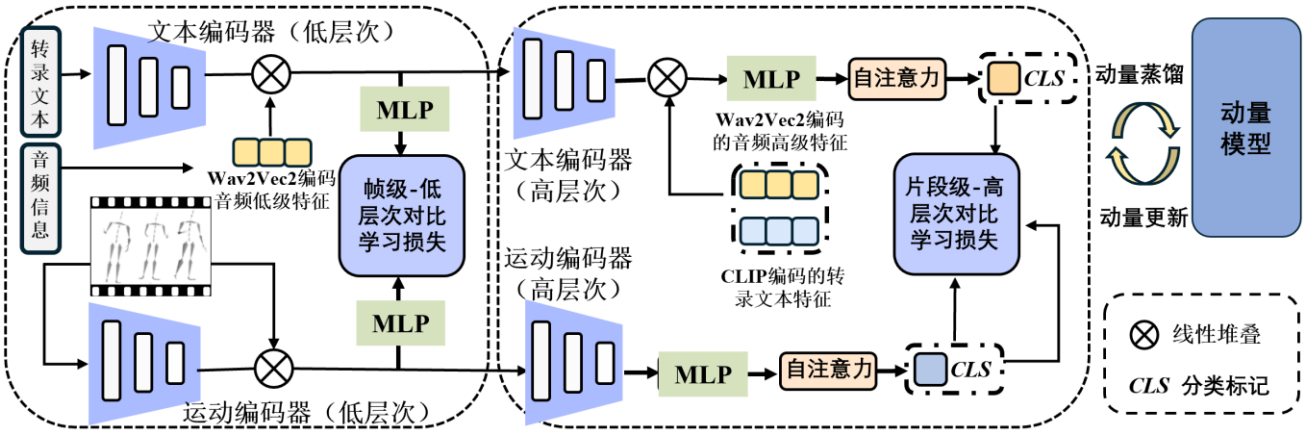

图 3 结合动量蒸馏模型的分层对比语义检索方法结构图

3.2 结合动量蒸馏的三维手势动作检索

3.2.1 三维动作检索框架结构

本文的手势数据中同时存在低级信号（如音频波形与手势关节坐标），高级语义（如文本含义与手势意图）的层级结构，需通过不同层级的检索分别建模，而单一检索层级难以同时满足细粒度时序匹配与整体语义一致的跨模态检索目标，若仅用片段级检索直接匹配句子语义-手势片段，容易出现局部时序异步性，导致语义正确但动作帧错位问题；若仅用帧级检索，可能丢失高级语义关联，导致动作帧对齐正确但语义出现错误。

基于此，本文提出了分层对比语义检索机制，并结合动量蒸馏的音频转录文本与三维动作进行对比检索。核心思想是多层次跨模态表示学习，两层级通过帧级提升片段级质量的联动逻辑，局部帧对齐为整体片段语义匹配提供基础，实现时序局部细节与高级语义的双重精准检索。

如图 3 所示，本文检索模型划分为帧级检索与片段级语义检索两个层级，本文编码器采用了 TANGO 模型[25]类似的双塔编码器架构，低层次文本特征编码器由 BERT[26]的前 6 层构成，BERT 的后 6 层结构连接 1 层 Transformer 层作为高层次文本编码器。运动表征编码器使用了 TM2T[27]的编码器结构，低层次的运动编码器由 10 层卷积神经网络构成，高层次的运动编码器由 18 层卷积神经网络与 1 层 Transformer 网络构成。

为增强模型表征能力，本文还集成了 Wav2Vec2[28]音频编码器以提升检索质量。帧级检索结合 Wav2Vec2 编码器中的卷积神经网络部分捕获声学细节特征，通过结合音频纹理来进行匹配。片段级语义检索通过 Wav2Vec2 中 Transformer 层提取音素级表征，并结合 CLIP[29]文本编码器以提供文本语义信息以实现多模态检索。

其中 Wav2Vec2 编码器主要通过对音频信号的局部时域、频域分析，主要捕捉音高、语速、停顿与音量等核心声学细节特征。这是由于人类对话中，手势的节奏、强度、时序与语音的声学特性天然存在同步关联，缺少这些特征会导致文本语义匹配但动作与语音动态脱节，增加 Wav2Vec2 捕捉的声学细节特征可辅助文本语义实现细粒度匹配，避免同一文本对应不同节奏的手势，而 CLIP 文本编码器定位语义，确保检索到的手势与用户意图一致。

帧级检索和片段级检索分离的架构目的是通过低层次信息捕获局部时域特征，再通过高层次信息建立更精确的语义检索能力，以实现从低阶信号到高阶表征的渐进式特征学习。需要注意的是，局部帧级别的检索目的是提升片段级语义检索质量，并不与后续扩散模型融合。

动量蒸馏模型的分层对比语义检索方法流程如图 3 所示。首先，针对文本模态的转录文本，其先经过 BERT 的前 6 层网络进行低层次特征编码，并集成 Wav2Vec2 编码的音频信息以提取特征；运动模态编码基于 TM2T 模型架构，手势信息通过低层次运动编码器捕捉局部帧级运动特征。在帧级检索阶段，文本-音频低层特征与运动低层特征通过余弦相似度匹配局部时域关联。片段级检索阶段，文本模态的转录文本通过高层次编码，结合 Wav2Vec2 编码的高级特征与 CLIP 编码的文本特征，实现语义深度整合，最终输出能够表征音频文本整体含义的 CLS（Classification Embedding）嵌入[26]；运动信息则经过高层次运动编码器，借助 18 层卷积与 1 层 Transformer 网络建模片段级运动依赖关系，生成聚合运动序列全局信息的 CLS 嵌入。在片段级语义检索阶段，各模态 CLS 嵌入用于计算全局语义相似度。其中，CLS 嵌入作为各模态特征的全局压缩载体，能够聚合单模态输入的整体语义与结构信息，避免局部特征碎片化导致的匹配偏差，确保多模态特征空间的一致性与检索精度。

最终，主模型依据上述流程生成各模态 CLS 嵌入与相似度分布，教师模型通过指数移动平均动

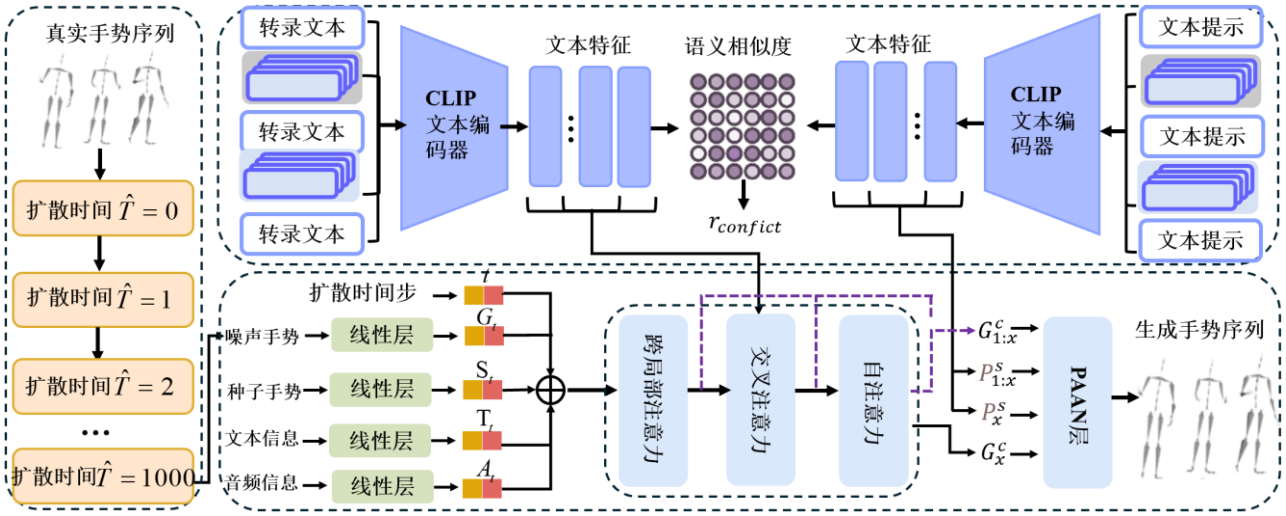

图 4 多模态信息多级注意力引导的扩散模型手势生成方法结构图

态更新参数以积累历史知识，其输出的 CLS 嵌入软标签通过 KL 散度损失约束主模型分布，有效缓解跨模态标注噪声导致的检索干扰。

3.2.2 三维动作检索训练策略与损失函数

本研究基于对比学习框架，其基础训练目标为 InfoNCE。在训练策略层面，模型同时采用全局对比学习损失与局部跨模态对比学习损失。其中设 B 表示批次，s 表示余弦相似度，a^{cls} 表示文本与音频的 CLS 嵌入，m^{cls} 表示动作的 CLS 嵌入，全局对比学习损失可表示为：

$$\mathrm{L}_{seg}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(s(a_i^{cls},m_i^{cls})\right)}{\sum_{j=1}^{B}\exp\left(s(a_i^{cls},m_j^{cls})\right)} \tag{2}$$

其中局部对比任务旨在构建具有时序敏感性的正负样本对，以当前帧为中心的时间邻近窗口 δ 帧内样本作为正例，其他邻近窗口 δ' 帧内的样本为负例。该设计旨在适应自然对话中语音与手势的时序异步特征。若 H 表示帧数的时间步数，k 表示当前时间帧，其中正样本时间偏移 $\delta\in\{-5,5\}$，P 表示正样本集合，N_{neg} 表示负样本时间偏移($\delta'\in\{-25,25\}$)，a^{low} 表示为音频与文本的低层特征，m^{low} 表示为运动的低层特征，局部对比学习的训练损失可表示为：

$$\mathrm{L}_{frame}=-\frac{1}{H}\sum_{k=1}^{H}\frac{1}{|\mathrm{P}|}\sum_{p\in\mathrm{P}}\log\frac{\exp\left(s(a_k^{low},m_{k+\delta}^{low})\right)}{\sum_{\delta'\in\mathrm{N}_{neg}}\exp\left(s(a_k^{low},m_{k+\delta'}^{low})\right)} \tag{3}$$

最终，InfoNCE 表示为：

$$\mathrm{L}_{InfoNCE}=\lambda\mathrm{L}_{seg}+(1-\lambda)\mathrm{L}_{frame} \tag{4}$$

需要注意的是，动量蒸馏的设计初衷，是在不脱离基础对齐目标的前提下，解决跨模态噪声干扰问题。若 L_{MoD} 的子损失仅保留 KL 散度，则蒸馏过程会脱离跨模态特征对比的核心任务，教师模型的软标签监督将失去基础对齐基准，导致模型仅拟合教师分布却忽视真实数据的模态关联，最终降低检索精度。所以 L_{MoD} 的子损失函数中出现 L_{frame} 和 L_{seg} 并非重复计算，而是使其约束在跨模态对比对齐的核心任务框架内，避免在动量蒸馏过程偏离核心任务。

3.2.3 结合动量蒸馏的三维动作检索

在音频-文本-手势的跨模态学习场景中，由于文本描述可能包含与手势动作无关的词汇，动作序列种也有未被文本描述的冗余手势。除此以外，构建的音频-文本对与真实手势运动也有可能存在语义偏差，这种模态间表征差异导致传统对比学习面临严重干扰，负样本可能意外匹配目标动作内容，而正样本亦难以保证语义一致性。

为了解决以上问题，本文受 ALBEF[30]等多模态对比学习框架启发，结合了动量蒸馏的鲁棒学习策略，构建由主模型与动量教师模型组成的双模型系统，本质是通过多轮学习的知识平滑，提升模型对模态噪声的泛化能力，理论上保证了即使数据存在局部噪声，仍能学到全局一致的跨模态关联。其中教师模型的参数 θ_{tea} 由指数移动平均算法动态更新，该机制使得动量模型能够积累层次检索模型的历史知识，生成稳定的伪标签作为辅助监督信号，动量模型为帧层级与片段层级生成软标签，通过 KL 散度损失迫使主模型的相似度分布逼近动量模型的分布，提升鲁棒性。令 q_{tea} 和 q_{stu} 分别为教师模型和主模型的相似度分布，L_{MoD} 可表示为：

$$\begin{aligned}\mathrm{L}_{frame}^{\mathrm{mod}}&=\lambda_1\mathrm{L}_{frame}\\&+(1-\lambda_1)KL(q_{tea}(a^{low},m^{low})\|\,q_{stu}(a^{low},m^{low}))\end{aligned} \tag{5}$$

$$\mathrm{L}_{\mathrm{seg}}^{\mathrm{mod}}=\lambda_2\mathrm{L}_{\mathrm{seg}} +(1-\lambda_2)KL(q_{\mathrm{tea}}(a^{\mathrm{cls}},m^{\mathrm{cls}})\| q_{\mathrm{stu}}(a^{\mathrm{cls}},m^{\mathrm{cls}})) \tag{6}$$

$$\mathrm{L}_{MoD}=\lambda_3\mathrm{L}_{frame}^{\mathrm{mod}}+(1-\lambda_3)\mathrm{L}_{seg}^{\mathrm{mod}}. \tag{7}$$

这种双重监督机制允许模型生成与原始标注存在合理偏差但语义相符的预测结果，并通过教师模型的缓变参数特性，有效缓解标注噪声带来的训练不稳定问题，总体检索模型损失函数可表示为：

$$\mathrm{L}_{total}=\lambda_4\mathrm{L}_{InfoNCE}+(1-\lambda_4)\mathrm{L}_{MoD}. \tag{8}$$

最终本研究通过索引标识符（数值编码+语义标题，如"1_FINGERS THREE"'）返回检索结果。需要注意的是，人类对话真实情况中，语义手势通常会在对应词汇出现前 0.4 秒左右开始预规划，并且在本文的数据集中，手势序列的 1/3 处，往往是手势能量最集中、语义表达最明确的阶段，因此本文会将检索到的时间位置与手势序列的三分之一的位置对齐，以实现手势先于匹配的关键字出现。

3.3 基于扩散模型的手势动作生成

3.3.1 扩散模型框架结构

扩散模型的框架如图 4 所示，本文通过扩散过程 q 获得符合正态分布的噪声手势。本文手势序列表示为 $G=[g_0,...,g_N]$，扩散过程会逐渐将高斯噪声添加到实际数据中，直到其分布接近于 $\mathrm{N}(0,\mathbf{I})$。具体来说，模型会按照预定义的时间 $\beta_t\in(0,1)$，逐渐向原始数据中添加高斯噪声 $g_0\sim p(g_0)$。最终在 $\hat{T}$ 步骤中将数据分布转化为各向同性的高斯分布，计算公式如下：

$$q(g_t|\,g_{t-1})=\mathrm{N}\left(g_t;\sqrt{1-\beta_t}g_{t-1},\beta_t\mathbf{I}\right). \tag{9}$$

去噪模块根据输入的手势噪声 g_t、去噪步长 t、音频 A_t、文本 T_t 和种子手势 S_t 重建原始信号 g_0：

$$\hat{g}_0=Denoise(g_t,t,A_t,T_t,S_t) \tag{10}$$

在去噪阶段，模型经过训练能够逆转扩散过程。这种训练使其在推理阶段能够将随机噪声转化为真实的数据分布。去噪过程的具体实现方式可表示为：

$$p_\theta(g_{t-1}|\,g_t)=\mathrm{N}\left(g_{t-1};\mu_\theta(g_t,t),\Sigma_\theta(g_t,t)\right). \tag{11}$$

假设 $\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$ 并且 $\alpha_t=1-\beta_t$，则时间 t 时的噪声手势 g_t 可以被表示为：

$$q(g_t|\,g_0)=\mathrm{N}\left(g_t;\sqrt{\bar{\alpha}_t}g_0,(1-\bar{\alpha}_t)\mathbf{I}\right). \tag{12}$$

接下来，本文考虑到手势生成任务对损失函数的要求，如精准拟合关节角度的连续值、抵抗数据中的异常姿态以及保障时序动作的平滑性，选择通过优化生成的手势 g_0 与真实人类手势 $\hat{g}_0$ 之间的 Huber 损失[31]来训练去噪模块。使用 Huber 损失主要考虑通过分段函数实现精度与稳健性的平衡：当误差小于阈值时，采用 MSE 损失[32]的二次函数形式，确保对正常样本的关节角度进行精准拟合，保留动作细节；当误差大于阈值时，切换为线性函数形式，避免异常值产生的极端误差过度干扰模型训练，从根本上减少异常样本导致的动作失真，这种特性较为匹配手势数据大部分正常姿态与少量异常值的分布规律。Huber 损失可以表示为：

$$\mathrm{L}_\delta(g_0,\hat{g}_0)=\begin{cases}\frac{1}{2}(g_0-\hat{g}_0)^2, & |g_0-\hat{g}_0|\le\delta\\ \delta(|g_0-\hat{g}_0|-\frac{1}{2}\delta), & |g_0-\hat{g}_0|>\delta\end{cases} \tag{13}$$

本节将扩散模型损失定义为 L_{net}，表示为：

$$\mathrm{L}_{net}=E_{g_0\sim q(g_0|\,c),t_d\sim[1,\hat{T}]}\left[\mathrm{L}_\delta\left(g_0,\hat{g}_0\right)\right] \tag{14}$$

3.3.2 扩散模型中的注意力机制

扩散模型生成手势的核心需求是利用多模态条件精准控制生成结果，而单一注意力机制难以覆盖局部动作细节、语义响应、全局时序连贯的多维度需求，因此本节扩散模型设计了三级注意力机制，由跨局部注意力，语义感知交叉注意力，自注意力构成。跨局部注意力在局部帧内计算注意力机制，核心解决局部细节问题。语义交叉注意力解决语义精准响应问题，以实现文本语义到手势特征映射。自注意力对局部注意力输出的特征进行全局时序建模，核心解决长序列连贯性与自然性的需求。

其中，跨局部注意力在传统注意力机制基础上，通过引入路由变换器[33]的局部特征关联策略，采用多通道聚合输入增强结构化信息流动。其计算形式为：

$$\mathrm{Attention}_{Local}=\mathrm{sofmax}\left(\frac{QK^{\mathrm{T}}+\bar{M}}{\sqrt{C_t}}\right)\cdot V, \tag{15}$$

Q、K、V 分别表示输入标记的查询、键、值矩阵，$\mathcal{T}$ 为矩阵转置运算符，$\bar{M}$ 为掩码矩阵，C_t 为通道维度。

由于人类的手势从来不是孤立的，它与头部姿态、躯干转动和腿部移动共同构成了完整的、富有表现力的肢体语言。为了确保生成结果在物理和认知上的合理性与真实性，本文通过语义交叉注意力

机制，将文本的高层语义信息注入到全身骨架的每个关节点的生成过程中。语义感知交叉注意力机制将跨局部注意力层输出的查询矩阵 Q 为输入，将 CLIP 编码的转录音频文本特征作为键值 K_{clip} 和值 V_{clip}，通过以下公式实现语义增强:

$$\text{Attention}_{semantic}=\text{softmax}\left(\frac{QK_{clip}^{T}}{\sqrt{C_t}}\right)\cdot V_{clip}. \quad (16)$$

此外，为了更好的融合手势运动的全局信息，最后本文还采用了自注意力机制，捕捉语句间的长期依赖，其中 Q、K、V 均源自语义注意力的输出特征，全局注意力架构能有效提升手势生成的语义相关性与运动连贯性。本文三级注意力机制通过局部细节、语义响应、全局连贯的递进逻辑，实现多模态条件的全粒度利用，此外，这种分层嵌入的设计也有利于提升模型的可解释性与可优化性，从而整体上提升了模型的控制精度和生成质量，最终提供更好的可控性，确保生成手势细节逼真、语义准确、手势连贯的效果。

3.3.3 扩散模型中的风格控制

现有生成模型的风格控制存在两大问题：一是风格注入生硬，易导致动作抖动；二是风格冲突失控（如音频语义“悲伤”但文本提示“激动”时，模型出现风格失效或者动作伪影），且风格系数多依赖手动调节，灵活性极差。本文提出 PAAN 模块解决风格注入生硬的问题，并结合扩散模型插值与无分类器引导进行缓解风格冲突问题。

首先为了提高系统的可控性，本文将局部注意力、语义注意力与自注意力输出定义为三层嵌入特征。输出沿通道维度串联，表示为 $G_{1:x}^{c}$，$P_{1:x}^{c}$ 表示由 CLIP 文本编码器编码的文本命令，其中，$1:x$ 是指多层特征的范围，x 仅表示最后一层。将 CLIP 编码的嵌入 $P_{1:x}^{s}$ 进行处理，使其与 $G_{1:x}^{c}$ 对齐，P_{x}^{s} 与注意力模块输出的 G_{x}^{c} 对齐。

如图 5 所示，本文确保运动和风格特征的维度一致性，本文对 $G_{1:x}^{c}$ 进行了以下操作：沿通道维度的均方差归一化、1×1 卷积和 N 帧跨度的平均池化。

对于 $P_{1:x}^{s}$ 和 P_{x}^{s}，本文通过 MLP 和池化操作进行处理。平均池化的目的是保证 N 帧之间的样式一致性，同时降低计算复杂度。在此基础上，本文提出了 PAAN(Prompt Adaptive Attention Normalization)

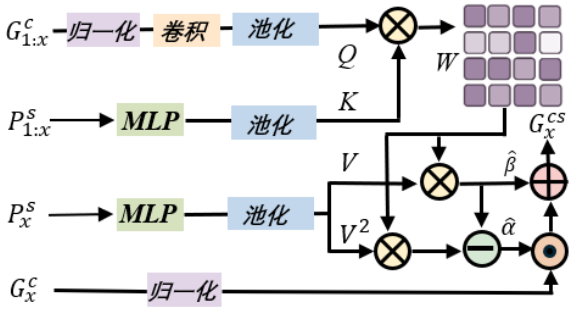

图 5 PAAN 网络结构

模块，输出的嵌入特征如下：

$$G_{x}^{cs}=PAAN(G_{x}^{c},P_{x}^{s},G_{1:x}^{c},P_{1:x}^{s}). \quad (17)$$

为了解决突出的局部风格特征导致的运动不稳定性问题，本文使用余弦相似性避免过度强调局部特征，提高运动一致性。使用余弦相似度来计算注意力权重 W 可表示为：

$$S_{i,j}=\frac{\mathrm{Q_i}\cdot\mathrm{K_j}}{|\mathrm{Q_i}|\times|\mathrm{K_j}|}+1 \quad (18)$$

$$W_{i,j}=\frac{\mathrm{S_{i,j}}}{\sum_{i}\mathrm{S_{i,j}}} \quad (19)$$

其中，元素之间的注意力权重 W 是通过对所有元素的相似度值进行归一化计算得出的。 根据注意力权重 W 可计算加权平均注意力 β 和加权标准差注意力 α，其定义如下：

$$\beta=W^{\bullet}\otimes V \quad (20)$$

$$\alpha=\sqrt{W^{\bullet}\otimes(V\cdot V)-\beta\cdot\beta} \quad (21)$$

最终转换后的特征图将归一化内容特征图与比例尺 α 和偏移量 β 进行加和，最终具有风格的手势 G_{x}^{cs} 可表示为：

$$G_{x}^{cs}=\alpha\cdot\mathrm{Norm}(G_{x}^{c})+\beta \quad (22)$$

3.3.4 扩散模型自动控制提示风格冲突

对于用户文本输入命令与音频语义出现不匹配甚至矛盾时（例如音频语义与转录文本都表示为激动的手势，但文本命令是悲伤难过的），这将同时影响手势的语义表达与风格提示的效果，本文将其定义为风格冲突问题。

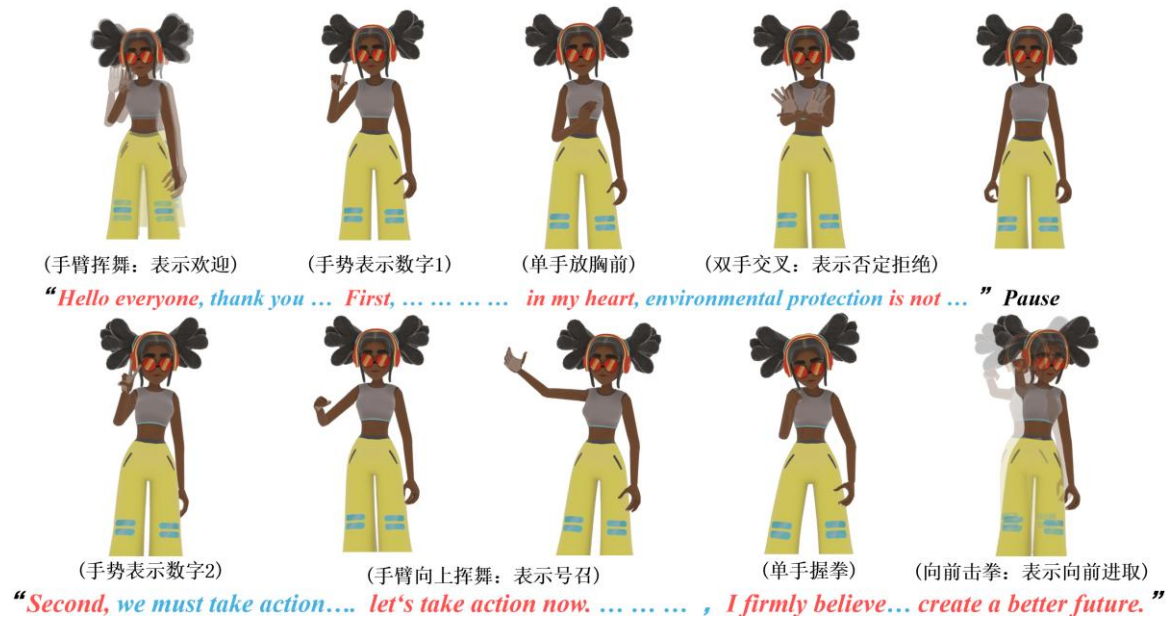

图 6 跨模态检索手势可视化分析

为了减轻过度的风格变化以及冲突问题导致动作不自然的问题，本文结合扩散模型插值与无分类器引导进行缓解，即通过使用参数 γ 在两个生成结果之间进行插值，从而生成风格受控的手势，可表示为

$$\hat{\varepsilon}\left(g_{1:\hat{T}}\right)=(1-\gamma)\hat{\varepsilon}\left(g_{1:\hat{T}},t,c,P\right)+\gamma\hat{\varepsilon}\left(g_{1:\hat{T}},t,c,\varnothing\right) \tag{23}$$

相关工作[18][20]中参数 γ 均采用手动控制，这影响方法的灵活性，因此本文提出一种自动控制风格的方法，通过替换关键词与引入否定词实现提示风格的反转。具体而言，首先，利用 NLTK 词性标注工具[34]提取关键风格相关形容词（例如“放松的”）。随后，借助 WordNet 工具[35]将其替换为反义词（例如将“放松的”替换为“紧张的”）。对于没有直接词典反义词的术语，本文采用添加“非”或“不”英文前缀，或在原始描述词前插入“不”等句子级英文否定运算符。对于已包含否定运算符的句子，系统直接通过依存解析工具 SpaCy 库[36]执行否定剥离，这种基于规则的方法在保持语义连贯性的同时，确保了句子间的提示反转。再通过测量转录文本 T 与 CLIP 编码嵌入的反转提示 $\bar{P}$ 之间的余弦相似度，同时使用了 SimCSE 模型[37]进行相似性计算，以 D_{SIMCSE} 表示。CLIP 编码结果和 SimCSE 模型的余弦距离加权和可表示为如下公式：

$$\gamma=\omega\left(1-\cos\left(\mathrm{D}_{\mathrm{CLIP}}(T),\mathrm{D}_{\mathrm{CLIP}}(\bar{P})\right)\right)+(1-\omega)\mathrm{D}_{\mathrm{SIMCSE}}(T,\bar{P}). \tag{24}$$

如公式（24）所示，音频与反转提示之间的相似度越高，表明冲突越强烈，这将导致 γ 值增大，继而公式（23）通过无分类器引自动降低风格注入强度，以减少抖动问题并最大限度减少不自然的语音手势。

3.4 基于掩码重采样的自然过渡方法

由于检索模块输出的语义手势与生成模块输出的手势来源独立，易出现姿态跳变，如前一帧手臂在胸口，后一帧突然到头顶等问题，违背人类手势的运动连续性规律，所以需要对多段动作进行衔接。目前主流方法基本通过数学拟合[38]构建过渡曲线实现衔接， 虽简单普遍，但存在显著缺陷：一是受人体关节动力学约束，插值生成的过渡帧易出现关节角度超限、运动轨迹违背生物规律的问题；二是脱离输入语义与音频节奏引导，若前后动作语义不同，会产生语义断裂，生成模糊动作。为了保证动作的真实性，本文提出基于掩码重新采样的方法进行动作衔接。本文首先将动作片段后缀与下一个动作片段前缀逐帧加权平均，其中 $\vec{a}$ 表示动态权重系数（ $\vec{a}\in[0,1]$），$\otimes$ 表示逐帧加权混合操作，S_i 表示当前运动序列，h 表示帧数，初步生成的运动序列 M' 可表示为：

$$M'=(1-\vec{a})\otimes S_{i-1}[-h:]+\vec{a}\otimes S_i[:h] \tag{25}$$

由于初步生成运动序列过渡存在生成不自然问题，本研究在基础上添加线性加权噪声，再通过去噪步骤将最初生成的运动进行精细化，令 M_{hard} 表示 0/1 值域的二值化遮罩， M_{soft} 表示连续值域的渐变遮罩，M'_{noisy} 表示添加噪声的运动序列，$*$ 表示

图 7 本文方法与其他方法的可视化比较

逐元素乘法操作，最终优化后的运动序列M''可表示为：

$$M''=M'+M_{\text{hard}}*M_{\text{soft}}*\left(M'_{\text{noisy}}-M'\right). \quad (26)$$

本文方法采用时序掩码矩阵对动作序列进行划分，确保已知区域的特征在扩散过程中保持不变，为过渡帧提供了明确的运动学锚点。前后有效帧的关节角度、运动速度等参数构成边界约束条件，使生成的过渡帧必须满足起始关节角度等于前序最后一帧、终止关节角度等于后序第一帧，并通过本文已训练好的三层注意力机制的去噪模型对混合区域重新进行生成，本质是通过扩散模型的后验优化，逐步去噪将噪声映射到符合真实数据分布的样本，优化过渡区动作的分布，这样利用扩散模型解决运动轨迹违背生物规律的问题，通过掩码已知区域和待生成区域方法避免运动轨迹的断裂问题，最终实现动作衔接之间的平滑过渡。

4 实验结果与分析

4.1 数据集与实验设置

4.1.1 数据集

本文检索模型使用 MOCCA 语义手势数据集[22]，MOCCA 数据集均为离散语义单元，其手势单元结构便于提取关键帧和语义标签，可以支撑分层对比检索的训练与查询，本文通过镜像运动（“左”,“右”）进一步增强，最终包含演员 1088 个 .bvh 格式的的动作捕捉数据，覆盖了常用的 208 种语义手势，每个手势平均以 2.6 种不同方式表示。本文对 MOCCA 数据集进行批量重定向，使其与 BEAT 数据集[39]的骨架对齐，最终动作库的骨架包含 75 个关节，以欧拉角 xyz 旋转，时间速率采样为 30fps。但需明确的是，本文使用的 MOCCA 数据集，其价值核心在于语义单元的结构化，而非数据量大小，且相比纯生成模型依赖的大规模无结构数据，其数据量显著更低，本质是通过高质量结构化降低对更多无结构化大数据量的依赖。

本文扩散生成模型使用 BEAT 数据集，该数据集包含大规模连续对话数据（约 76 小时），提供了丰富的韵律与常见手势动作，适合训练生成模型学习常见的语义动作与节奏同步的手势，本文将原始数据重采样为 30 FPS，以欧拉角 xyz 旋转，使用了英语标注的说话者进行训练，其中包括两名说话者（一男一女），测试集包括另外三名说话者（两男一女）。

4.1.2 实验设置与训练过程

在检索模型架构中，参数λ,λ_4均设置为 0.8，λ_1,λ_2,λ_3分别设置为 0.7，0.7，0.2。检索模型使用 256 个批次进行训练，在单个 NVIDIA A800 GPU 上训练了大约 2 天。扩散模型生成架构中，跨局部注意机制使用 8 个头和 48 个注意力通道，窗口大小为 15 帧。交叉注意机制与自注意力机制由 8 层 8 个头组成。参数 D_{clip}、C_{ada}、ω，N 分别设置为 512、384、0.2，10。扩散模型使用 128 个批次进行训练，扩散模型框架在单个 NVIDIA A100 GPU 上训练了大约 4 天。

4.2 评估指标与用户研究设计

为了衡量本文中生成手势的语义质量，本文采用了以下主要指标：(1) Frechet Gesture Distance（FGD）Error! Reference source not found.利用预训练的自动编码器将动作投射到潜在空间，用来评估的生成输出与真实分布之间差异。(2) Diversity(Div)[40]通过

测量生成手势的变化来评估多样性。(3) Semantic-Relevance Gesture Recall（SRGR）[39]是 BEAT 数据集提出基于时间语义权重正确匹配比例的加权版本。

此外，本文通过人类主观评估观察生成的结果，通过 SoJump 系统对生成的手势序列与数据集中的真实动作数据进行了用户研究。本研究将这些指标定义为人类相似度、语义适当性、风格化程度和风格适当性。这项研究有 25 人参与，参与者均来自本校，年龄在 20 至 30 岁之间，所评估的短片长度从 30 秒到 60 秒不等。为了避免偏差，本文调整了视频顺序，参与者在以 1 分至 5 分之间进行打分。

4.3 对比实验

4.3.1 跨模态检索可视化对比

如图 6 所示，本研究将检索手势动作进行了可视化，当出现“Hello”等表示欢迎的动作时，手势表现为挥舞手臂；当出现“First”、“Second”等动作时，本文方法用手指表达对应的数字；当出现“not”等动作时，双手交叉表示否定。这些语义内容与检索到的手势动作合理匹配，这证明了检索模型的有效性。

4.3.2 生成手势的可视化对比

如图 7 所示，本文方法与 DiffSHEG Error! Reference source not found.、DiffGesture Error! Reference source not found.、DiffuseStyleGesture+ [41]和 CaMN[39]等方法进行比较，同时可视化了没有使用本文提出的检索模型，而是仅使用扩散模型进行手势生成的效果。以图 7 说话内容为例，当音频内容出现“Amazing”时，本文方法表现出竖大拇指的动作，这与其他四种方法相比明显更加自然。此外，若仅使用本文提出的扩散模型，虽然也能做出了类似惊叹赞扬的手臂外展动作，但与本文整体流程中检索到的语义手势有一定表现力差距，这说明将语义检索机制与扩散生成模型结合能拥有更好的语义性与表现力。

4.3.3 风格控制效果对比

如图 8 所示，本研究直观地展示了 LivelySpeaker[42]、GestureDiffuCLIP Error! Reference source not found. 与本文方法的对比结果，可以观察到风格命令对于生成结果的影响。本文方法在文本提示为“激动”，音频情感风格为“悲伤、疲惫”时，依然能自动控制风格效果，不会出现剧烈抖动与伪影问题。这并非依赖特定标注数据训练达成，而是模型自主捕捉到语义与动作的通用关联规律，本文通过语义注意力机制与 PAAN 层进一步强化了这风格的精准性。它会根据文本语义的风格倾向，动态调整全身骨架特征的

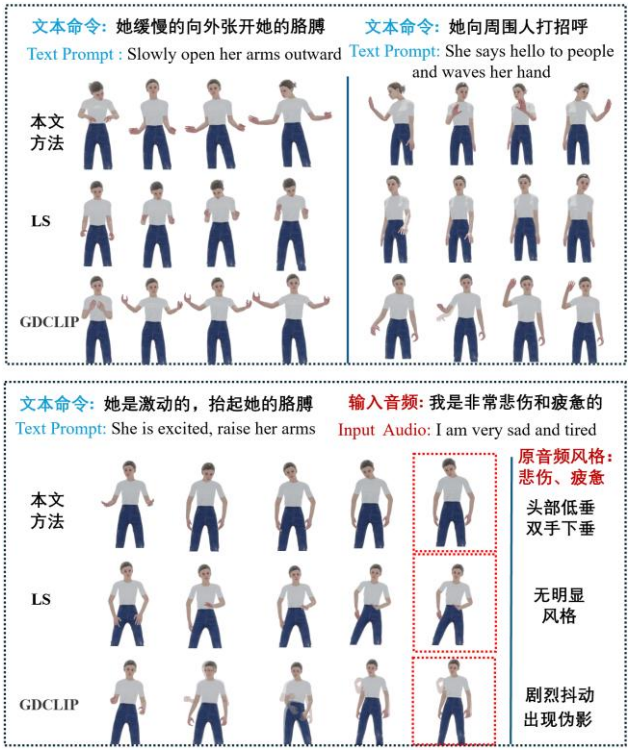

图 8 风格提示效果对比

统计分布，对比 LivelySpeaker[42]方法直接采用 CLIP 编码器编码文本提示的方法，这导致了风格效果不够明显。GestureDiffuCLIP Error! Reference source not found. 采用将文本提示命令注入到 AdaIN 层进行风格注入的方式，虽然表现出了一定的风格效果，但出现剧烈抖动与伪影问题。尤其当提示风格与输入音频发生冲突，抖动现象会变得更加明显。

4.3.4 分层检索模型性能对比

本研究通过检索运动序列的正确匹配率来评估模型性能，并通过验证结果是否高于随机检索来评估模型的有效性。其中评估体系包含帧级别特征检索与片段级别语义检索两个维度，如表 1 所示：

表 1 分层对比语义检索机制的评估对比

系统	帧级别检索	片段级别检索
随机检索	4.23%	2.31%
硬编码关键字检索	23.67%	35.67%
CLIP 编码的语义检索	10.56%	55.17%
本文方法	**75.38%**	**87.28%**

在帧级别特征检索中，评估方法设计如下：随机选取第 i 帧音频特征，在运动特征序列的[i-25, i+50]帧窗口内计算余弦相似度。若最大相似度对应的运动帧位于[i-5, i+5]区间，则判定为有效匹配。基于 5000 次随机采样的统计结果显示，该任务的随机搜索理论准确率为 4.23%，而本文方法达到

75.38%的检索精度，显著超越基准水平。

片段级别语义检索任务采用 1:1087 的正负样本比例构建评估集，具体流程为：给定音频，在包含 1087 个干扰项的候选池中计算跨模态相似度。若最大相似度对应真实配对运动样本，则视为有效匹配。在 1000 次独立测试中，系统取得 87.28%的检索准确率，相比随机搜索提升显著。

本文还与硬编码关键字检索，使用 CLIP 编码器编码文本后与动作库描述进行对比检索的方式，进行对比。实验结果表明，本文方法指标评分明显优于上述两方案，证明了本文检索机制的有效性。

4.3.5 手势生成模型性能对比

本文模型还与 CaMN[39]、DiffSHEG **Error! Reference source not found.**、DiffuseStyleGesture+[41]、GestureDiffuCLIP**Error! Reference source not found.**、DiffGesture**Error! Reference source not found.**、LivelySpeaker[42]等方法，以及本文方法流程中不使用检索模型仅使用扩散模型的方式，进行定量指标对比与用户评估。

表 2 本文方法定量指标对比与用户评估

系统	定量指标			用户评估	
	FGD ↓	Div ↑	SRGR ↑	人类相似度↑	语义适当性↑
真值	0	11.52	1	$4.18^{\pm 0.40}$	$4.38^{\pm 0.50}$
CAMN[39]	8.653	8.44	0.208	$3.44^{\pm 0.51}$	$3.75^{\pm 0.44}$
DSG+[41]	10.877	10.69	0.203	$3.06^{\pm 0.44}$	$4.13^{\pm 0.34}$
DSHG**Error! Reference source not found.**	9.176	10.11	0.195	$3.44^{\pm 0.32}$	$3.53^{\pm 0.34}$
DG**Error! Reference source not found.**	15.356	9.89	0.116	$2.03^{\pm 0.18}$	$2.13^{\pm 0.19}$
LS[42]	9.234	9.27	0.184	$3.50^{\pm 0.52}$	$3.19^{\pm 0.66}$
GDC**Error! Reference source not found.**	9.615	9.89	0.106	$3.36^{\pm 0.18}$	$3.43^{\pm 0.19}$
本文扩散模型方法	**8.456**	10.23	**0.219**	$4.06^{\pm 0.57}$	$4.18^{\pm 0.54}$
本文方法	9.160	**12.77**	0.205	$\mathbf{4.37}^{\pm \mathbf{0.51}}$	$\mathbf{4.55}^{\pm \mathbf{0.19}}$

如表 2 所示，本文提出的扩散模型生成手势在 FGD、SRGR 指标获得最高得分，但本文提出的整体流程方法并未在这些指标超过仅使用扩散模型。这是由于 FGD、SRGR 定量指标是用来衡量生成手势与 BEAT 数据集手势特征相似程度，而本文的整体流程方法并不完全依赖于 BEAT 数据集，还会检索到比 BEAT 数据集中手势更具语义的手势动作，这点可以由可视化结果与用户评估获得最高得分观察到。此外，本文方法在多样性指标也获得了很高的分数，这证明了与其他方法相比，本文方法拥有更强的手势多样性与语义表现力。

本文还对比不同损失函数在手势扩散生成模型训练中的性能。实验保持模型的网络结构，仅替换损失函数，并从三个定量核心指标开展评估。

结果如表 3 可见，采用 Huber 损失时：在动作真实度、多样性与语义匹配度上均优于 MSE 损失，Huber 损失在保证正常样本拟合精度的同时，避免了异常值的过拟合，削弱了异常值对语义动作映射关系的干扰，使生成手势更精准响应输入语义信息。

表 3 扩散生成模型不同损失函数对指标的影响

系统	FGD ↓	Div ↑	SRGR ↑
Huber 损失	8.456	10.23	0.219
MSE 损失	9.548	8.38	0.183

此外，本文还评估了方法整体的效率表现，模型对于一段约 1 分钟的音频输入（对应 900 帧手势序列），在单张 NVIDIA GeForce RTX 4060 GPU 上，完整流程平均推理时间仅为 96.7 秒。这显著优于仅依赖本文扩散模型（平均 135.4 秒），检索机制在成功匹配场景下带来了约 28.6%的速度提升，优化效果直接归功于检索阶段对高质量语义手势片段的高效选取（平均成功检索到 4.2 个以上语义手势片段），有效减少了后续扩散模型的帧数比例，从而降低了整体计算开销。

本文与生成效率最快的 DiffSHEG**Error! Reference source not found.**方法进行了同硬件环境下的对比，DiffSHEG **Error! Reference source not found.**原始 DDPM 方案推理耗时高达 1944.1 秒。其优化策略采用了 25 步 DDIM 采样和提出的 FOPPAS 方法，虽然将平均耗时大幅降至 42.6 秒，但这种加速是以显著牺牲生成手势的质量为代价的。相比之下，我们的方法在保持高质量、高语义匹配度和丰富细节的前提下，达到了更具实际应用价值的推理速度。

4.3.6 本文风格控制方法用户评估对比

本文方法还与 LivelySpeaker[42] (LS)、GestureDiffuCLIP (GDC) **Error! Reference source not found.**方法进行了风格控制的比较。同时，本研究还通过手动控制了风格系数 γ，与本文自动控制风格强度的方式进行比较。

如表 4 所示，本文方法在除风格化程度外，均

优于 LivelySpeaker[42]、GestureDiffuCLIP Error! Reference source not found.与手动确定风格参数。

表 4 本文风格控制方法与其他方法用户评估对比

系统	人类相似度↑	语义适当性↑	风格化程度↑	风格适当性↑
LS [42]	3.50 ±0.52	3.19 ±0.66	3.42 ±0.51	3.56 ±0.26
GDC Error! Reference source not found.	3.67 ±0.34	3.44 ±0.36	3.72 ±0.43	2.96 ±0.37
手动控制参数-0.5	4.05 ±0.33	3.85 ±0.54	3.52 ±0.42	3.98 ±0.44
手动控制参数-1	3.96 ±0.25	3.58 ±0.45	**4.08 ±0.34**	3.78 ±0.24
手动控制参数-3	1.88 ±0.34	2.23 ±0.34	2.85 ±0.57	2.33 ±0.52
本文方法	**4.18 ±0.57**	**4.11 ±0.42**	4.05 ±0.37	**4.09 ±0.54**

当风格系数设置为 1 时，风格化效果最佳，但风格系数应该根据情况调整，不能直接设置为固定数值。实验还可以观察到，增加风格系数评分反而下降，这是由于虽然较高的风格系数强度会使手势的风格更加明显，但过高的强度会导致不自然的手势和抖动问题。本文方法通过 PAAN 层以及自动控制风格系数，动态调整风格强度，不仅实现了更明显、更清晰的风格注入效果，还大大缓解了局部抖动与伪影问题。

4.4 消融实验

4.4.1 分层检索组件消融

本文对分层对比语义检索机制组件进行了消融实验，如表 5 所示：

表 5 分层对比语义检索机制不同组件对性能的影响

系统	帧级别检索	片段级别检索
本文方法	**75.38%**	**87.28%**
无 Wav2Vec2	70.11%	73.45%
无 CLIP	-	70.33%
无 Wav2Vec2&CLIP	-	66.45%
无动量蒸馏机制	69.84%	62.77%
无帧级别检索	-	60.87%

从表 5 可观察到，当去除 Wav2Vec2 或 CLIP 编码器后检索性能都会下降，同时去除后下降程度会更加显著。本文中的结合动量蒸馏方法为检索模型提高稳定性与噪声鲁棒性，帧级别的检索为模型提供了更多的信息，去除两者后，检索性能都会出现明显下降。这说明本文提出的各个组件对分层对比语义检索机制都具有重要意义。

4.4.2 扩散模型组件消融

本文对扩散模型的多个组件进行了消融实验，分别为：（1）使用 FastText 转录文本，（2）局部注意力机制，（3）交叉注意力机制，（4）自注意力机制，（5）PAAN。需要注意的是，在考察语义影响时，本文排除了 PAAN 层和风格相关指标。在考察风格效果时，本文去除 PAAN 层，并采用直接调整风格参数的方式。同时，本文也消融了多种组合组件引起的结果变化，以证明多种组件组合对定量指标和用户评估的影响。

表 6 评估了扩散生成模型不同组件对定量指标的影响。表 7 评估了扩散生成模型不同组件对用户评估的影响。从表 6、表 7 结果可以看出，当不使用跨局部注意力时，各项分数会持续降低，这说明局部特征为手势生成提供重要信息。当移除转录文本或交叉注意力时，会导致评分进一步下降，这说明了多模态手势数据具有重要作用。当不使用自注意力时，各项指标评分断崖式下降，这表明在语音-手势异步性的特殊条件下，关注手势全局信息比局部特征更加重要。

表 6 扩散生成模型不同组件对定量指标的影响

系统	FGD ↓	Div ↑	SRGR ↑
真值	0	11.52	1
无转录文本	12.753	9.77	0.195
无局部注意力	13.227	9.5	0.207
无交叉注意力	11.297	10.11	0.205
无自注意力	15.753	8.12	0.193
无局部&交叉注意力	15.343	9.22	0.111
无局部&自注意力	17.654	7.13	0.174
无交叉&自注意力	19.245	7.33	0.109
本文方法	**8.456**	**10.23**	**0.219**

表 7 扩散生成模型不同组件对用户评估的影响

系统	人类相似度↑	语义适当性↑	风格化程度↑	风格适当性↑
真值	4.18 ±0.40	4.38 ±0.50	-	-
无转录文本	3.65 ±0.62	3.81 ±0.63	-	-
无局部注意力	3.13 ±0.62	3.44 ±0.51	-	-
无交叉注意力	3.18 ±0.48	3.06 ±0.65	-	-
无自注意力	3.38 ±0.81	3.37 ±0.66	-	-
无局部&交叉注意力	2.85 ±0.64	2.66 ±0.43		
无局部&自注	3.03 ±0.30	2.95 ±0.72		

意力				
无交叉&自注意力	2.66 $^{\pm 0.44}$	2.33 $^{\pm 0.65}$		
无 PAAN 层	3.86 $^{\pm 0.71}$	3.93 $^{\pm 0.68}$	2.86 $^{\pm 0.80}$	2.56 $^{\pm 0.81}$
本文方法	**4.06** $^{\pm 0.57}$	**4.18** $^{\pm 0.54}$	**3.89** $^{\pm 0.27}$	**3.06** $^{\pm 0.22}$

在对多种组件组合进行消融时，消融自注意力机制会使 FGD 与 Div 指标出现显著下降，全局结构的破坏会直接降低动作的真实感。而只要消融交叉注意力会导致 SRGR 以及用户评估（如语义适当性）的下降更为明显。当交叉注意力被移除，直接造成生成手势与输入语义脱节，因此语义匹配指标 SRGR 与用户感知会出现显著滑坡，侧面证明了多模态语义融合的重要性。

在评估风格效果时，消融 PAAN 层采用直接拼接，而非 PAAN 层进行风格嵌入融合时，生成动作会出现明显抖动与不自然运动，所有评估维度的评分均大幅下降。这证明 PAAN 层能有效提升风格化效果的自然度，并显著缓解运动抖动问题。这些实验证明了，本文使用的各个模块对扩散模型生成高质量手势动作都存在重要作用。

4.5 局限性分析

尽管本研究在手势生成领域取得了一定进展，但仍存在一些局限性：首先，虽然 MOCCA 数据集虽然结构化程度高，能覆盖大多数情况手势表达情况，但覆盖的语义手势类型和人体形态仍可能存在不足。这导致模型在面对库外的罕见手势或特定人群的手势时，检索精度会下降，生成质量也会受影响，这是基于检索的方法共同面临的闭集问题。其次依赖检索模块的准确性，本文框架依赖检索模块的准确性。一旦检索模块因噪声或歧义返回了不恰当的手势片段，会大幅影响真实性和沉浸感，并且目前检索模块对音频的节奏特征利用不足，生成手势的节奏与语音匹配度仍有提升空间。

未来工作将考虑解决这些局限性,一方面会扩展手势库的语义覆盖范围和多样性，另一方面尝试引入在线学习与个性化适应，允许模型在部署后，通过与特定用户的交互，学习并更新其手势库，实现个性化手势生成。此外，未来工作将继续精细化检索与生成的全流程，搭建检索结果的智能验证与纠错机制，对检索到的手势片段进行匹配度量化评估并依据阈值取舍。在此基础上，进一步将检索手势的特征映射至语义-节奏联合高维空间，融入音频的语速、重音、停顿等节奏特征，以生成兼具精准语义表达与自然节奏韵律的手势动作。

5 结语

本文提出一种基于跨模态语义检索与条件扩散生成的三维手势协同合成框架。通过分层对比语义检索机制实现文本、音频与动作参数的多模态语义检索，并结合多级注意力引导的扩散模型生成自然且多样化的手势动作，最终通过掩码重采样生成过渡自然的三维手势动作。实验结果表明，本方法在弗雷歇手势距离、语义匹配度等指标上优于现有手势生成方法，生成结果在动作语义性表达、风格可控性方面展现出优势。

尽管本研究在手势生成领域取得了一定进展，但仍存在一些局限性：本文构建的多模态手势库虽然涵盖了多数语义手势场景，但手势类别覆盖范围仍有限，难以满足极端体型角色或复杂交互场景的需求。此外，当前架构主要关注检索模型的语义表现，涉及节奏变化较少，可能导致检索到的动作节奏变化表现力受限。未来将进一步扩充手势动作库的规模，同时关注具有音频节奏的检索手势动作表达，以覆盖更多更复杂手势交互场景。

参 考 文 献

[1] Studdert-Kennedy M. Hand and mind: what gestures reveal about thought. Language and Speech, 1994, 37(2): 203-209.

[2] Cooperrider K, Wakefield E, Goldin-Meadow S. More than meets the eye: Gesture changes thought, even without visual feedback //Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Pasadena, California, USA, 2015: 441-446.

[3] Galati A, Weisberg S M, Newcombe N S, Avraamides M N. When gestures show us the way: Co-thought gestures selectively facilitate navigation and spatial memory. Spatial Cognition and Computation, 2018, 18(1): 1-30.

[4] Friedman A, Cafaro F. From thoughts to interaction: designing controls for video playback gestures with embodied schemata//Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg, Germany, 2023: 146:1-146:7.

[5] Delamare W, Silpasuwanchai C, Sarcar S, Shiraki T, Ren X. On gesture combination: an exploration of a solution to augment gesture interaction //Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces. Daejeon, Republic of Korea, 2019: 135-146.

[6] Wagner P, Malisz Z, Kopp S. Gesture and speech in interaction: an overview. Speech Communication, 2014, 57: 209-232.

[7] Kopp S, Wachsmuth I. Model-based animation of co-verbal gesture //Proceedings of the Computer Animation 2002 (CA 2002). Geneva, Switzerland, 2002: 252-257.

[8] Neff M, Kipp M, Albrecht I, Seidel H-P. Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Transactions on Graphics, 2008, 27(5): 1-5:24.

[9] Habibie I, Elgharib M, Sarkar K, et al. A motion matching framework for controllable gesture synthesis from speech //Proceedings of the ACM SIGGRAPH 2022. Vancouver, Canada, 2022:46: 1-46:9.

[10] Yoon Y, Cha B, Lee J H, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics, 2020, 39: 1-16.

[11] Li Jing, Kang Di, Pei Wen-Jie, Zhe Xue-Fei, Zhang Ying, He Zhen-Yu, Bao Lin-Chao. Audio2gestures: generating diverse gestures from speech audio with conditional variational autoencoders //Proceedings of the IEEE International Conference on Computer Vision 2021. Montreal, Canada, 2021: 11293-11302.

[12] Alexanderson S, Henter G E, Kucherenko T, Beskow J. Style-controllable speech-driven gesture synthesis using normalising flows. Computer Graphics Forum, 2020, 39(2): 487-496.

[13] Yi H, Liang H, Liu Y, et al. Generating holistic 3D human motion from speech //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2023. Vancouver, Canada, 2023: 469-480.

[14] Rombach R, Blattmann A, et al. High-resolution image synthesis with latent diffusion models //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2021. Nashville, USA, 2021: 10674-10685.

[15] Chen J, Liu Y, Wang J, et al. DiffSHEG: A diffusion-based approach for real-time speech-driven holistic 3D expression and gesture generation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2024. Seattle, USA, 2024: 7352-7361.

[16] Ao Teng-Long, Zhang Ze-Yi, Liu Li-Bin. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. ACM Transactions on Graphics, 2023, 42(4):1-18.

[17] Zhu L, Liu X, Liu X, et al. Taming diffusion models for audio-driven co-speech gesture generation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023. Vancouver, Canada, 2023: 10544-10553.

[18] Yang Si-Cheng, Xu Zun-Nan, Xue Hai-Wei, et al. FreeTalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2024. Seoul, Republic of Korea, 2024: 7945-7949.

[19] Li L, Li W, Ding Q, et al.. Gesture generation via diffusion model with attention mechanism //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2024. Seoul, Republic of Korea, 2024: 8316-8320.

[20] Yang Si-Cheng, Wu Zhi-Yong, Li Ming-Lei, Zhang Zhen-Song, Hao Lei, Bao Wei-Hong, Cheng Ming, Xiao Long. DiffuseStyleGesture: Stylized audio-driven co-speech gesture generation with diffusion models //Proceedings of the International Joint Conference on Artificial Intelligence 2023. Macao, China, 2023: 5860-5868.

[21] Zheng Rui-Kun, Liu Geng-Xin, Hu Rui-Zhen. Semi-supervised Character Animation Style Transfer. Journal of Computer-Aided Design and Computer Graphics,2025,37(5): 821-831 (in Chinese)
(郑瑞坤, 刘耿欣, 胡瑞珍.半监督角色动画风格迁移.计算机辅助设计与图形学学报, 2025, 237(5): 821-831)

[22] Zhang Ze-Yi, Ao Teng-Long, Zhang Ying, et.al Semantic gesticulator: semantics-aware co-speech gesture synthesis. ACM Transactions on Graphics (TOG), 2024, 43(4): 1-17.

[23] Yang Z., Li L., Lin K., Wang J., Lin C. C., Liu Z., Wang L. The dawn of lmms: preliminary explorations with gpt-4v (ision). Ithaca, NY, USA: Cornell University Library (arXiv), technical report: arXiv:2309.17421, 2023.

[24] Tan Xu, Chen Jun, Liu Hong, Cong Jun, Zhang Chao, Liu Ying, Liu T Y. Naturalspeech: end-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4234-4245.

[25] Liu Hong, Yang Xin, Akiyama T, et al. Tango: co-speech gesture video reenactment with hierarchical audio motion embedding and diffusion interpolation. arXiv preprint arXiv:2410.04221, 2024.

[26] Devlin J, Chang M W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[27] Guo C, Zuo X, Wang S, et al. Tm2t: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts //Proceedings of the European Conference on Computer Vision 2022. Cham: Springer Nature Switzerland, 2022: 580-597.

[28] Babuškaš I, Baevski A, Conneau A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations //Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. Virtual, 2020: 12449–12460.

[29] Radford A, Baevski A, Conneau A, et al. Learning transferable visual models from natural language supervision //Proceedings of the 38th International Conference on Machine Learning 2021. Virtual, 2021: 8748–8763.

[30] Li Jun-Nan, Selvaraju R., Gotmare A., Joty S., Xiong Cai-Ming, Hoi S. C. H. Align before fuse: vision and language representation learning with momentum distillation //Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2021: 9694-9705.

[31] Huber P J. Robust estimation of a location parameter. Annals of Statistics, 1964, 53 (1): 73-101.

[32] Bishop C M. Pattern recognition and machine learning. Cham: Springer, 2006.

[33] Roy A, Saffar M, Vaswani A, et al. Efficient content-based sparse attention with routing transformers. Transactions of the Association for Computational Linguistics, 2021, 9: 53-68.

[34] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. Sebastopol: O'Reilly Media, 2009.

[35] Fellbaum C. WordNet: an electronic lexical database. Cambridge,USA: MIT Press, 1998.

[36] Honnibal M, Johnson M. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io, 2017

[37] Gao T, Yao X, Chen D. Simcse: simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821, 2021.

[38] Yang S, Wu Z, Li M, Zhang Z, Hao L, Bao W, Zhuang H. Qpgesture: quantization-based and phase-guided motion matching for natural speech-driven gesture generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 2321-2330.

[39] Liu H, Zhu Z, Iwamoto N, et al. Beat: a large-scale semantic and emotional multi-modal dataset for conversational gestures

synthesis//Proceedings of the European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 612-630.

[40] Bhattacharya U, Childs E, Rewkowski N, et al. Speech2affectivegestures: synthesizing co-speech gestures with generative adversarial affective expression learning//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China, 2021: 2027-2036.

[41] Yang S, Xue H, Zhang Z, et al. The DiffuseStyleGesture+ entry to the GENEA Challenge 2023//Proceedings of the 25th International Conference on Multimodal Interaction. Paris, France, 2023: 779-785.

[42] Zhi Y, Cun X, Chen X, et al. Livelyspeaker: towards semantic-aware co-speech gesture generation//Proceedings of the IEEE/CVF International Conference on Multimodal Interaction. Paris, France, 2023: 20807-20817.

WANG Xin-Yi, M.S. candidate. His research interests include virtual human gesture generation.

LIU Shi-Guang, Ph.D., professor. His research interests include computational graphics, computer animation and virtual reality.

YANG Zhao-Meng, undergraduate student. Her research interests include dance studies and dance education.